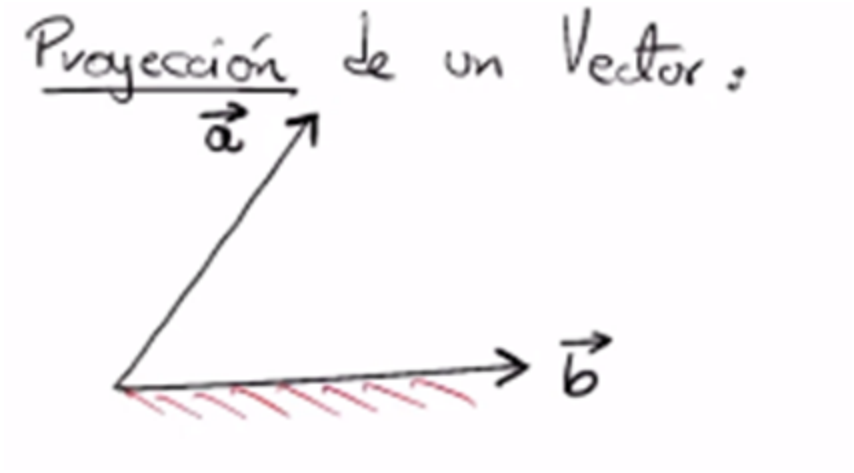
PCA: Análisis de componentes principales

Este tema trata obre como al aplicar procesamiento, codificación a variables categóricas resulta en tener más datos y que no necesariamente son relevantes, con esta herramienta lo que se trata de llevar a cabo es realizar un análisis para poder eligir adecuadamente que variables o parámetros se van a elegir para poder meter al modelo de Machine Learning.

De manera que si hay correlaciones interesantes, podemos reducir el numero de variables solo a las que nos arrojan la información minima necesaria, entonces para esto tenemos el concepto de correlación. Vamos a ver como esto y los elementos de Algebra Lineal nos ayudan para aplicar esta técnica.

Proyección de un vector

Tenemos 2 vectores $\vec{a}$ y $\vec{b}$ con direcciones contrarias y hablamos de la **Proyección (Sombra)** de un vector sobre otro geometricamente de la siguiente manera. Imaginemos que la sombra se va proyectar sobre el eje del vecor $\vec{b}$, así:



Y la sombra que se refleja del vector $\vec{a}$ sobre la superficie que es donde esta el vector $\vec{b}$, es esto:

Es la distancia que esta marcada en la linea roja, a esto se le llama **La proyección**

Y se define como una longitud nueva $\vec{a_p}$ entonces lo definimos como:

$$\vec{a_p} = a_p \, \hat{b} = \left(\frac{\vec{a} \cdot \vec{b}}{|\vec{b}|} \right) \hat{b}$$

Donde:

- $\vec{a_p} =$ Vector a_p
- $a_p =$ Longitud de a_p
- $\hat{b} =$ Vector unitario en b

Ejemplo

- $\vec{a} = (2, 2)$
- $\vec{b} = (1, 0)$

Calculemos los demás parámetros

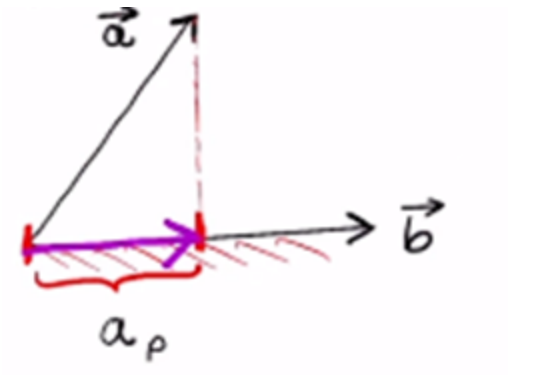
- $|\vec{b}| = \sqrt{(X_b)^2 + (Y_b)^2} = \sqrt{(1)^2 + (0)^2} = 1$
- $\hat{b} = (1, 0) = \vec{b}$
- $\vec{a} \cdot \vec{b} = a_x * b_x + a_y * b_y = 2 * 1 + 2 * 0 = 2$

Obteniendo $\vec{a_p}$:

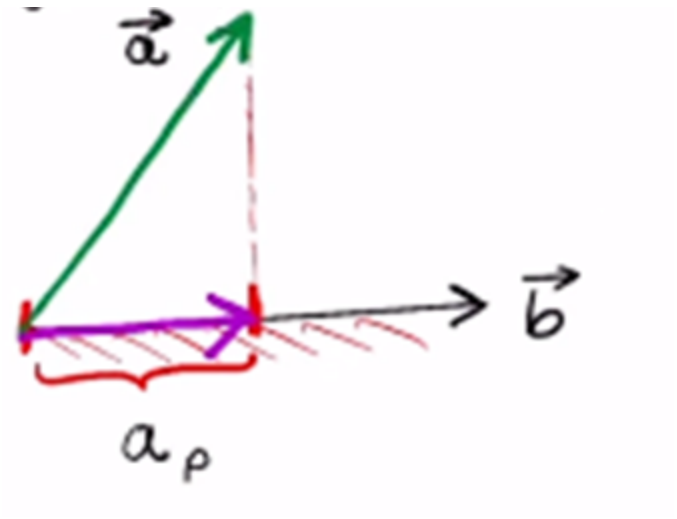
$$\vec{a_p} = \left(\frac{2}{1}\right) \hat{b}$$

$$\vec{a_p} = 2\hat{b} = 2(1, 0) = (2, 0)$$

Entonces $\vec{a_p}$ corresponde al vector en morado

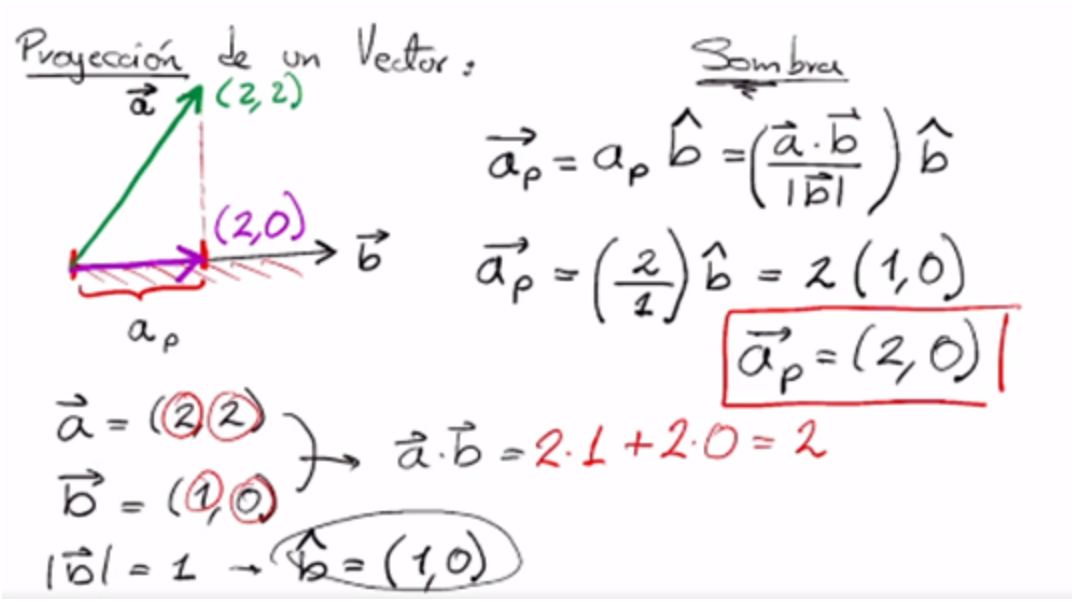


Mientras que el color original era el vector en color verde:



Lo que se hizo fue anular la componente en Y y se proyecto sobre el vector $\vec{b}$

Procedimiento completo:



¿Para qué nos va a servir el concepto de proyectar un vector?

Resulta que así como tenemos la matriz de co-varianza, nosotros podemos hablar de cuál es la varianza a lo largo de una cierta dirección en un conjunto de datos ¿cómo así?.

Respuesta

Resulta que cuando tienes un conjunto con varias variables, cada conjunto tiene una columna de datos y cada uno de esos datos tiene una distribución asociada.

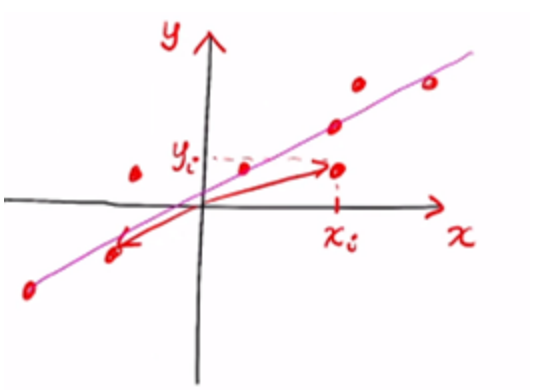
Imagina un plano de 2 dimensiones, un gráfico de dispersión; tenemos 2 variables {La que va en el eje x y la que va en el eje y} y tenemos toda la dispersion de todos los puntos. Entonces ahí nosotros podríamos decir que tanto están variando los datos si me muevo en x o si me muevo en y, o si me muevo en cualquier dirección arbitraria. De esto resulta que habrá ciertas direcciones en donde yo capturo mejor la varianza de los datos.

Veamoslo geometricamente:

Varianza en una dirección arbitraria

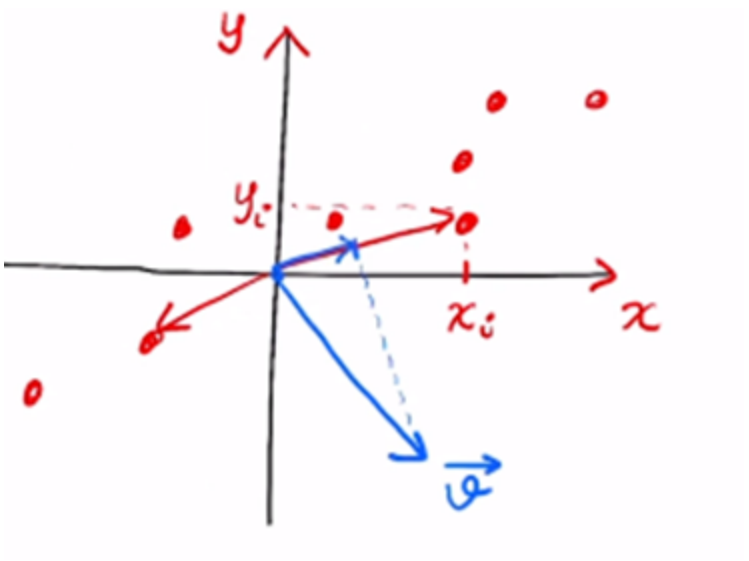
Esta gráfica representa un conjunto de datos respecto a x e y. De manera que cada uno de los puntos tiene su coordenada en x_i e y_i.

Aquí podemos ver claramente que puede haber una especie de correlación, a lo largo de la dirección marcada en color morado.

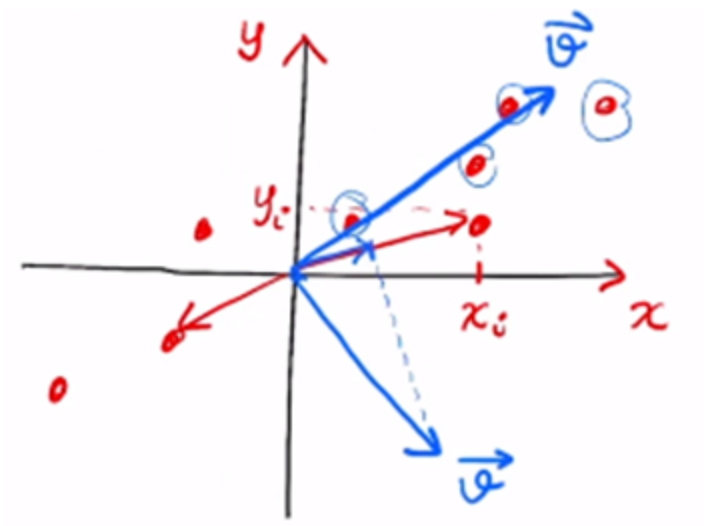


Entonces veamos como entra en juego el concepto de proyectar un vector.

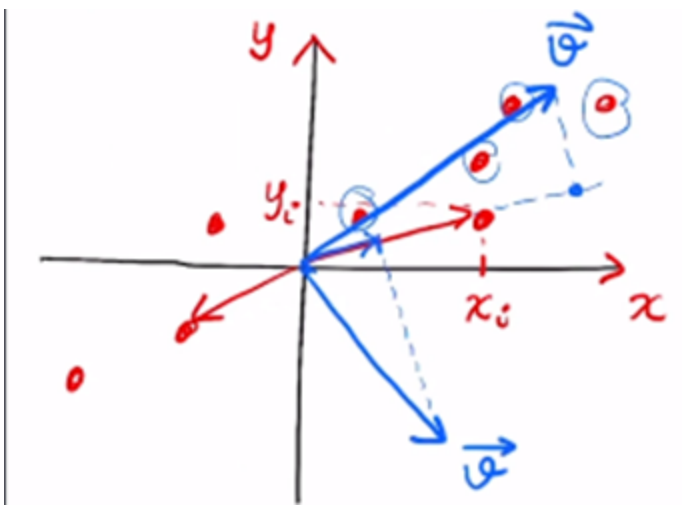
Yo voy a imaginarme una dirección cualquiera en el espacio (digamos una dirección dibujada por el vector que esta en azul) lo llamaré $\vec{v}$ y lo que veo cuando aplico el concepto de proyección es la superposición de un vector en azul, sobre el vector rojo:



Entonces se puede hacer en todos los conjuntos de datos o puntos. Lo que vemos en general es si yo cambio la dirección del vector $\vec{v}$ a una que este mas cerca a los datos. Es decir otro vector que este en esta dirección:

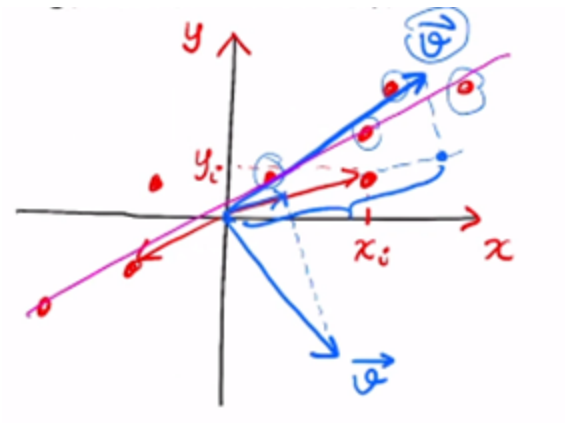


Entonces la sombra que el proyecta sobre el vector rojo ahora será la extensión que le dibujamos:

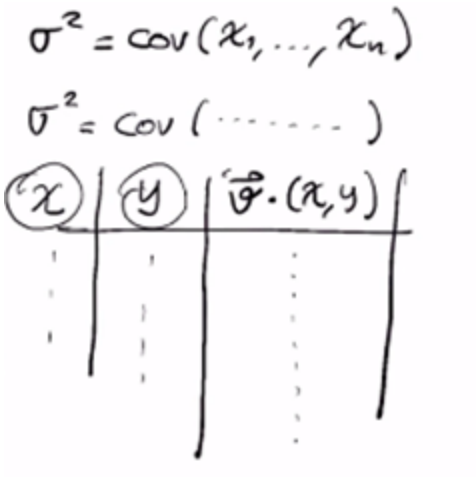


Y lo que vemos en general es que la longitud de la sombra es mayor; es decir cuando la dirección del vector que yo estoy usando va mas en relación con la dirección que tienen los datos (gráfica morada), las proyecciones son mayores, como se ve en la gráfica.

> Nota; Por otro lado en el caso anterior era diferente, porque no iba en dirección a los datos.



Y entonces ahí decimos que la varianza que se esta capturando en esa dirección (dirección de los datos) es mayor. ¿Cómo se define eso?
Si antes teníamos que la $covarianza$ era $\sigma^2 = cov(X_1, \ldots, X_n)$
Ahora decimos que la $covarianza$ sera la proyección del vector $\vec{v}$ sobre los puntos del set de datos, $\sigma^2 = cov(\ldots\ldots\ldots)$



Lo que significa que no se calcula la convarianza con las variables originales sino que lo calculo con relación a $\vec{v}$ y (x, y) yo puedo tener una covarianza asociada, no a los datos como tal sino a una dirección de movimiento que yo estoy tomando, de esta manera esto lo podemos escribir así

$$var(\vec{v} \cdot \vec{x_1}, \ldots, \vec{v} \cdot \vec{x_n})$$

Esto es interesante porque el termino anterior se puede escribir en términos de la **matriz de covarianza** como sigue:

$$var(\vec{v} \cdot \vec{x_1}, \ldots, \vec{v} \cdot \vec{x_n}) = \vec{v}\, \Sigma\, \vec{v}^T$$

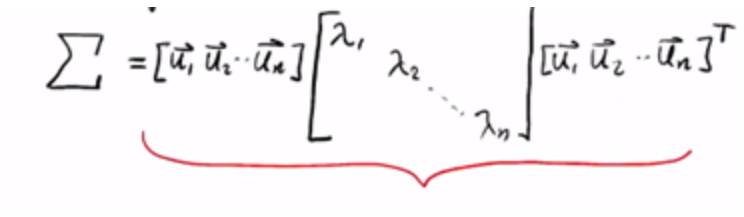
Podemos checar esta formula en un libro de Algebra Lineal

Recapitulando

Lo que nos damos cuenta es que la varianza en una cierta dirección es tomar la **matriz de covarianza** Σ y multiplicarla con el vector de esa dirección $\vec{v}$. Lo que estamos viendo en la anterior definición

$$\vec{v}\, \Sigma\, \vec{v}^T$$

es que necesitamos calcular **Valores y vectores propios**



¿Y qué es lo que sucede?

Aquí lo interesante era ver era la analogía de la manera en que como calculo la varianza en un vector y la manera en que como puedo descomponer en si misma la matriz de covarianza de un conjunto de datos. Porque en Álgebra Lineal se sabe y ese es otro resultado matemático que lo que son los valores propios y los vectores propios caracterizan completamente la varianza de un conjunto de datos en términos de lo que se llaman **LAS COMPONENTES PRINCIPALES** ¿CUALES SON LAS COMPONENTES PRINCIPALES? Son los vectores y valores propios que están en la ecuación. Al calcular todo lo aplicaremos a Python y después veremos que significa

Extras

- PCA Análisis
- PCA